Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)
http://dare.uva.nl/document/114027

File ID	114027
Filename	Chapter 5: Experimental evaluation

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type	Dissertation
Title	People search in the enterprise
Author	K. Balog
Faculty	Faculty of Science
Year	2008
Pages	IX, 196
ISBN	978-90-9023247-8

FULL BIBLIOGRAPHIC DETAILS:
http://dare.uva.nl/record/281835

Copyright

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other then for strictly personal, individual use.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)

5

Experimental Evaluation

We have detailed our expertise retrieval models in Chapter 3, and discussed our experimental setup in Chapter 4. In this chapter we present an experimental evaluation of our models. The research questions we seek to answer in this chapter concern the comparison of Model 1 and 2 (RQ 1/1), the choice of window sizes for Model 1B and 2B (RQ 1/2), a comparison of the baseline and window-based models, that is Model 1 vs. 1B and 2 vs. 2B (RQ 1/3), and the generalizability of models across different collections (RQ 7). In two largely self-contained sections, we describe the outcomes of experiments for the expert finding (Section 5.1) and expert profiling (Section 5.2) tasks. We summarize our findings in Section 5.3. In this chapter we focus on the main findings; a more fine-grained analysis and topic-level examination of the results is postponed until Chapter 6.

5.1 Expert Finding

We evaluate our expert finding models on the 2005–2007 editions of the TREC Enterprise test sets. The 2005 and 2006 editions use the W3C document collection (Section 4.4.1), while the 2007 edition uses the CSIRO (Section 4.4.2) document collection. We report on the measures listed in Section 4.3 and in all cases we use the boolean document-candidate association method (Eq. 3.15).

Next, we present the outcomes of our experiments. One by one, we address the research questions listed in Section 4.1.

5.1.1 Baseline Models: Model 1 vs. Model 2

Which of Model 1 and Model 2 is more effective for finding experts? (RQ 1/1) The results of the comparison are presented in Table 5.1.

Several things are worth noting. Concerning the 2005 and 2006 topic sets we find that the scores achieved on the 2006 collection are high, in absolute terms, for all measures. In fact, they are substantially higher than the 2005 scores; this is most likely due to the differences in assessment procedure used (see Section 4.4.1). More-

Model	TREC 2005		TREC 2006		TREC 2007	
	MAP	MRR	MAP	MRR	MAP	MRR
1	.1883	.4692	.3206	.7264	.3700	.5303
2	**.2053**	**.6088**$^{(2)}$	**.4660**$^{(3)}$	**.9354**$^{(3)}$	**.4137**$^{(1)}$	**.5666**

Table 5.1: Model 1 vs. Model 2 on the expert finding task, using the TREC 2005–2007 test collections. Best scores for each year are in boldface.

over, on the 2006 collection Model 2 clearly outperforms Model 1, on all measures, and these differences are statistically significant. On the 2005 collection, the picture is more subtle: Model 2 outperforms Model 1 in terms of MAP and MRR; however, the difference in MAP scores is not significant.

As to the 2007 topic set, the MAP scores are in the same range as those of TREC 2006 and the MRR scores are in the same range as those of TREC 2005. Again, Model 2 outperforms Model 1, but the gap between the two is narrower than on the 2005 and 2006 topics. The difference in MAP is significant, though.

In conclusion, Model 2 outperforms Model 1 on all topic sets, significantly so on the 2006 and 2007 topic sets in terms of MAP and on the 2005 and 2006 topic sets in terms of MRR. It is important to point out that on the 2006 test set, where the ground truth is more lenient (human generated), all differences are highly significant.

5.1.2 Window-based Models: Models 1B and 2B

Next, we look for performance differences between models based on different window sizes, i.e., for Models 1B and 2B (RQ 1/2). Recall that for Models 1B and 2B the candidate-term co-occurrence is calculated for a given window size w, after which a weighted sum over various window sizes is taken (see Eq. 3.17). Here, we consider only the simplest case: a single window with size w, thus $W = \{w\}$ and $p(w) = 1$.

To be able to compare the models, first the optimal window sizes (for MAP and MRR) are empirically selected for each model and topic set. The range considered is $w = 15, 25, 50, 75, 100, 125, 150, 200, 250, 300$.[1] The MAP and MRR scores corresponding to each window size w are displayed in Figure 5.1; notice that the ranges on the y-axis are different for MAP and MRR.

According to the plots on the left-hand side of Figure 5.1, in terms of MAP the ideal window size is between 75 and 250, and the MAP scores show small variance within this range. Model 1B on the TREC 2005 topic set seems to break this pattern of behavior, and delivers its best performance in terms of MAP at window size 25. In terms of MRR, however, smaller window sizes tend to perform better on the 2005 and 2006 topics; this is not suprising, as smaller windows are more likely to generate high-precision co-occurrences. For TREC 2007 the ideal window size in terms of MRR seems to coincide with that for MAP, hence is in the range of 75–200.

[1] Here we follow Cao *et al.* (2006), who consider window sizes 20, ..., 250; note that the average document length is approximately 500 words for W3C and 350 words for CSIRO.

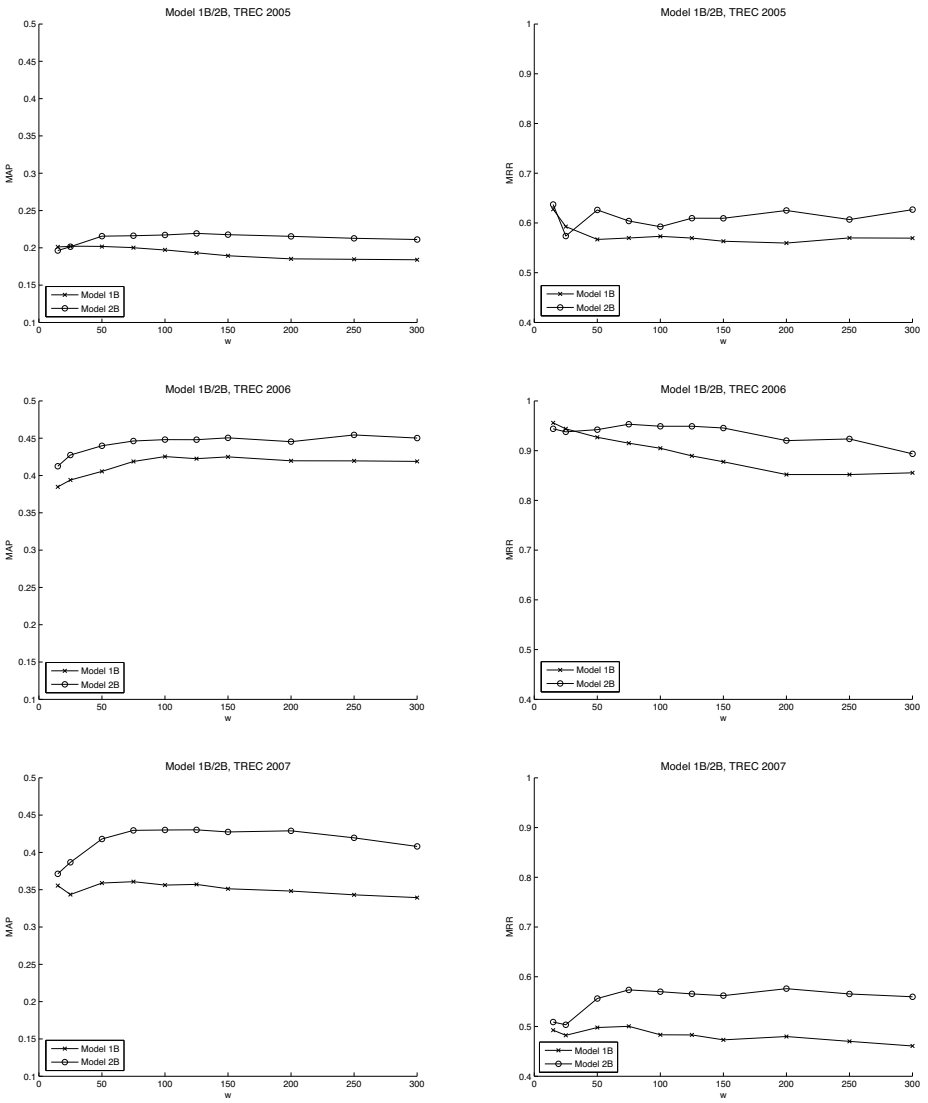

Figure 5.1: Effect on MAP (Left) and MRR (Right) of varying the window size w. Results are validated on the 2005–2007 TREC topic sets (Top–Bottom).

It is worth pointing out that the difference between the best-performing and worst-performing window size is statistically significant on the 2005 and 2006 topic sets for both measures (MAP and MRR) and both models (1B and 2B). On the 2007 topics this only holds for Model 2B and MRR.

5.1.3 Baseline vs. Window-based Models

What is the effect of lifting the conditional independence assumption between the query and the candidate? That is, what if any, are the performance differences between the baseline models (Model 1 and Model 2) and the window-based models (Model 1B and Model 2B, respectively) (RQ 1/3)? For the window-based models, we use the best performing configuration, i.e., window size, according to the results of the previous subsection. We present two sets of results, one based on window sizes optimized for MAP (Table 5.2), and one based on window sizes optimized for MRR (Table 5.3).

Model	TREC 2005			TREC 2006			TREC 2007		
	w	MAP	MRR	w	MAP	MRR	w	MAP	MRR
1	–	.1883	.4692	–	.3206	.7264	–	.3700	.5303
1B	25	.2020	.5928$^{(1)}$	100	.4254$^{(3)}$	.9048$^{(3)}$	75	.3608	.5003
2	–	.2053	.6088	–	**.4660**	**.9354**	–	.4137	**.5666**
2B	125	**.2194**	**.6096**	250	.4544	.9235	125	**.4303**	.5656

Table 5.2: Overall results on the expert finding task; window sizes optimized for MAP. Best scores (per measure) for each year are in boldface.

Looking at the results in Table 5.2 we find that, for the MAP-optimized setting, Model 1B outperforms Model 1 on the W3C collection (TREC 2005 and 2006 topic sets). On the TREC 2005 topics, the improvement is most noticeable in early precision: MRR +26% vs. MAP +7%; the difference in MRR is significant. On the TREC 2006 topics the improvement of Model 1B over Model 1 is even more substantial, achieving +32% MAP and +24% MRR; the differences in both MAP and MRR are highly significant. In contrast, Model 1B actually performs worse than Model 1 on the CSIRO collection (TREC 2007 topic set), but the differences are not significant.

As to Model 2B, it delivers improvements in MAP on the TREC 2005 and 2007 topics, but is outperformed by the baseline (Model 2) on the 2006 topics. On the other hand, MRR slightly drops on the 2006 and 2007 topics. Nevertheless, none of the differences between Model 2 and Model 2B are significant (i.e., neither for MAP, MRR, nor for 2005, 2006, or 2007).

Finally, Model 2B performs better than Model 1B, but the gap between them is smaller than between Model 2 and 1. The differences between Model 1B and 2B are significant (at confidence level 0.999) in MAP on the 2007 topics, but not for any other measure/topic set.

Next we turn to a comparison between the baseline and window-based models based on MRR-optimized settings; see Table 5.3. Model 1B improves over Model 1 on the 2005 and 2006 topics, and there the improvement is significant in all but one case (2005, MAP). Comparing Models 2 and 2B we observe a slight improvement in MRR on all topic sets, but there are losses in MAP on 2005 and 2006; none of the differences are significant. Finally, as to the differences between Model 1B

Model	TREC 2005			TREC 2006			TREC 2007		
	w	MAP	MRR	w	MAP	MRR	w	MAP	MRR
1	–	.1883	.4692	–	.3206	.7264	–	.3700	.5303
1B	15	.2012	.6275$^{(2)}$	15	.3848$^{(1)}$	**.9558**$^{(3)}$	75	.3608	.5003
2	–	**.2053**	.6088	–	**.4660**	.9354	–	.4137	.5666
2B	15	.1964	**.6371**	75	.4463	.9531	200	**.4289**	**.5760**

Table 5.3: Overall results on the expert finding task; window sizes optimized for MRR. Best scores (per measure) for each year are in boldface.

and Model 2B, the following are significant: 2006 MAP$^{(2)}$, 2007 MAP$^{(2)}$, and 2007 MRR$^{(1)}$.

5.1.4 Summary

In this section we evaluated our expert finding models on the 2005–2007 editions of the TREC Enterprise test sets. We found that Model 2 outperformed Model 1 on all topic sets and measures, in most cases significantly so.

Further, we investigated the window-based variations of our baseline models: Model 1B and 2B. We focused on the simplest setting, where a single window size (w) is considered (instead of a mixture of weighted window sizes). An empirical exploration over a range of windows sizes (15–300) showed that the optimal value of w varies across topic set, model, and measure combinations. In most cases the difference between the best-performing and worst-performing window size is statistically significant. Overall, apart from a few outliers, best MAP scores are achieved using a window size in the range of 75–125, and best MRR scores in the range 15–75.

Concerning the comparison of our baseline models against their window-based variations (Model 1 vs. 1B and Model 2 vs. 2B), we witnessed improvements in a number of cases. However, neither Model 1B nor 2B managed to outperform the corresponding baseline model on all topic sets. When compared to each other, Model 1B and 2B display the same relative behavior as Model 1 and 2, namely, Model 2B outperforms Model 1B. On the other hand, the gap between them is smaller than between Model 2 and 1, but (with one exception) the differences are not significant.

5.2 Expert Profiling

Now we change tack and we evaluate the performance of our baseline models on the expert profiling task. For experimental evaluation, we use the UvT collection (Section 4.4.3); this collection naturally fits the profiling task at hand as ground truth was obtained from an interface where people self-assessed their skills against a given set of knowledge areas (from now onwards: topics).

Given the nature of the UvT document types[2] and the fact that we have explicit authorship information for each document, we argue that going below the document level—i.e., taking the proximity between candidate occurrences and terms into account—is likely to have little or no impact. Therefore, we limit our evaluation to Model 1 and 2 (RQ 1/1), and do not address research questions that concern the window-based models (RQ 1/2 and RQ 1/3).

5.2.1 Model 1 vs. Model 2

How, then, do Model 1 and 2 compare on the expert profiling task (RQ 1/1)? Table 5.4 presents the results. We can see that Model 2 outperforms Model 1 on both languages and metrics. All differences are highly significant. The models deliver very similar results across languages.

Language	UvT ALL			
	Model 1		Model 2	
	MAP	MRR	MAP	MRR
English	.2023	.3913	**.2682**$^{(3)}$	**.4968**$^{(3)}$
Dutch	.2081	.4130	**.2503**$^{(3)}$	**.4963**$^{(3)}$

Table 5.4: Model 1 vs. Model 2 on the expert profiling task, using the UvT test collection. Best scores for each language are in boldface.

5.2.2 All Topics vs. Main Topics

Given that the self-assessments are sparse in our collection, in order to get a more reliable measure of performance, we selected a subset of topics, referred to as *UvT MAIN*. This set consists of 136 topics that are located at the top level of the topical hierarchy (see Section 4.4.3). In other words, a main topic has subtopics, but is not a subtopic of any other topics. The relevance judgments were also restricted to the main topic set, but were expanded with assessments of subtopics; see Section 4.4.3

[2]The UvT collection contains four types of documents: research descriptions, course descriptions, publications, and personal homepages (see Section 4.4.3). Research and course descriptions are very short and focused, and they are very unlikely to contain person references. Personal homepages are in a very heterogeneous format, and contain a limited number of name occurrences (other than those of the authors); other names are mostly found in publication lists. Publications do contain person references, or rather, references to other publications. Authors of these referred publications could be extracted from the bibliography and then mapped to Tilburg University employees (where appropriate). Developing such an extraction mechanism, however, would require a non-trivial effort, while the expected benefits are very limited. Recall that our interest is limited to creating profiles of UvT employee and to this end we seek "person-word" co-occurrences; hence, we only care about citations of publications by UvT employees. But, if an author that is being cited is indeed from Tilburg University, the publication is likely to exist in our document collection, and hence we would harvest person-word occurrences there. Therefore, this special treatment of publications would be of benefit only in a handful of cases, in particular, where a Tilburg employee cites another Tilburg employee and the full-text version of the cited publication is not available.

for details. Table 5.5 shows the performance of Model 1 and 2 (using English and Dutch topic translations) on the two UvT topic sets: ALL and MAIN.

Language	UvT ALL				UvT MAIN			
	Model 1		Model 2		Model 1		Model 2	
	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR
English	.2023	.3913	**.2682**$^{(3)}$	**.4968**$^{(3)}$	.3003	.4375	**.3549**$^{(3)}$	**.5198**$^{(3)}$
Dutch	.2081	.4130	**.2503**$^{(3)}$	**.4963**$^{(3)}$	.2782	.4155	**.3102**$^{(3)}$	**.4854**$^{(3)}$

Table 5.5: Model 1 vs. Model 2 on ALL vs. MAIN topics of the UvT collection. Best scores for each language are in boldface.

Looking at Table 5.5 we see that Model 2 outperforms Model 1 on the MAIN topics, and the difference between the two models is approximately the same on both topic sets. Differences between Model 1 and 2 are highly significant for the MAIN topics as well. When we compare results across languages, we find that the difference between English and Dutch is more apparent on the MAIN topics than on ALL topics.

All Dutch Topics

Recall that in Section 4.4.3 we restricted ourselves to those topics (986), where both English and Dutch translations of the knowledge area were available. As a side remark, we report on all Dutch topics vs. those 986 "shared" Dutch topics (that is, UvT ALL, where both English and Dutch translations of the topic are available). Table 5.6 summarizes the results.

Topic set	# topics	Model 1		Model 2	
		MAP	MRR	MAP	MRR
UvT ALL (UK and NL)	986	.2081	.4130	**.2503**$^{(3)}$	**.4963**$^{(3)}$
Dutch ALL (NL only)	1,499	.1882	.4282	**.2239**$^{(3)}$	**.5006**$^{(3)}$

Table 5.6: Model 1 vs. Model 2 on the expert profiling task, using all Dutch topics.

Moving from the topic set UvT ALL to Dutch ALL, the number of knowledge areas considered for creating expertise profiles increases by more than 50%. Despite this major growth in the number of topics, and hence, a corresponding growth in sparseness, the MAP scores drop only by around 10%. On the other hand the MRR scores do not decrease, but even increase slightly. The explanation behind this behavior of MRR is that it is easier to select a single topic that fits a person's expertise very well from a larger pool of topics. Again, we see that Model 2 outperforms Model 1, significantly so. The difference between UvT ALL and Dutch ALL is only significant for Model 1, MRR at the 0.95 level.

These results suggest that our methods are robust, continuing to perform at comparable levels even when data sparseness is increased substantially.

5.3 Summary and Conclusions

In this chapter we presented an experimental evaluation of our models on the expert finding and expert profiling tasks. Going back to the research questions we set out for this chapter, we conclude that Model 2 is more effective than Model 1 for expertise retrieval—both for the expert finding and expert profiling tasks (RQ 1/1).

Further, we performed an empirical investigation of window sizes for Model 1B and 2B, and found that (i) the optimal window size varies across topic set, model, and measure combinations, and (ii) various window sizes lead to significant differences in performance (RQ 1/2). Lifting the conditional independence assumption between the query and the candidate (Model 1 vs. 1B and Model 2 vs. 2B) leads to improvements in a number of cases. However, neither Model 1B nor 2B is able to outperform the corresponding baseline model on all topic sets (RQ 1/3).

Our models displayed consistent behavior across collections and tasks, in particular, Model 2 outperformed Model 1 for all collections and topic sets. This leaves us with the conclusion that our models generalize well across collections (RQ 7).

The results we discussed in this chapter were based on averaged numbers over the whole topic set. In the next chapter we analyze our main findings by drilling down to the level of individual topics (Section 6.1). Moreover, in this chapter we used automatic parameter settings and did not address the issue to what extent performance depends on the choice of the models' (smoothing) parameters. This parameter sensitivity analysis is carried out in Section 6.2 below.